Seis vulnerabilidades de IA. Tres brechas. Una respuesta arquitectónica.

Una analista de cumplimiento lee una advertencia de vulnerabilidad. Salesforce acaba de corregir ForcedLeak: una inyección de prompt oculta en un campo de formulario Web-to-Lead que hacía que un agente de IA consultara el CRM y exfiltrara registros a un dominio que el atacante compró por cinco dólares. Ella revisa la implementación de Agentforce en su empresa. La IA interactúa con cientos de leads al día y tiene acceso completo al CRM. Pregunta al área de seguridad si existe algún control —independiente del modelo de IA— que evalúe cada solicitud de datos según la política antes de que la IA los recupere.

La respuesta es no. Esa brecha no se resuelve con un parche.

Seis divulgaciones. Tres brechas. Un año.

Entre mediados de 2025 y abril de 2026, investigadores revelaron seis vulnerabilidades críticas de IA en las plataformas que las empresas usan a diario. Todos los proveedores lanzaron parches. Todos los ataques explotaron brechas arquitectónicas que ningún parche puede cerrar.

Vulnerabilidad	Plataforma	Brecha(s)	Qué falló
EchoLeak (CVSS 9.3)	Microsoft 365 Copilot	1 + 2	Correo electrónico manipulado como contexto de IA; acceso implícito amplio; exfiltración a través de dominios confiables
ForcedLeak (CVSS 9.4)	Salesforce Agentforce	1 + 2	Campo de formulario como contexto de IA; acceso total al CRM; dominio en lista permitida expirado por $5
GeminiJack	Google Gemini Enterprise	1 + 2	Documento manipulado indexado por RAG; barrido sin clics por todos los datos de Workspace
Reprompt	Microsoft Copilot	1 + 2	Parámetro de URL como contexto de IA; exfiltración con un solo clic
GrafanaGhost	Grafana AI	1 + 3	Datos de registros de eventos como contexto de IA; proceso a nivel de sistema con alcance funcional excesivo. RBAC existe, pero nunca se activó
OpenAI Plugin	OpenAI Ecosystem	2 + 3	Plugin comprometido recolectó credenciales; 6 meses, 47 empresas

Estas seis presentan tres brechas distintas:

Brecha 1: Entrada no confiable como contexto confiable de IA. Todos los ataques comienzan con datos externos —un correo, un documento compartido, un campo de formulario, entradas de registros de eventos— que ingresan al sistema y son procesados por la IA sin validación. Presente en las seis. El sector prácticamente lo ignora.

Brecha 2: Acceso amplio de IA a datos sin control por operación. Cinco de seis implican IA que se autentica una vez y luego accede a todo lo que está a su alcance. Ninguna evaluación de políticas por operación limita cada recuperación.

Brecha 3: Fallos de contención de procesos. GrafanaGhost operó mediante procesos internos a nivel sistema – no en una sesión de usuario—. El proceso tenía un alcance funcional para el que nunca fue diseñado. El ataque al plugin de OpenAI tuvo éxito porque las credenciales estaban almacenadas en un lugar accesible para el código comprometido.

Las barreras a nivel de modelo fallaron en todos los casos. La de Grafana cayó ante una sola palabra clave. La CSP de Salesforce fue burlada por cinco dólares. Las barreras son configuraciones internas del sistema atacado. Complementan controles reales. No los sustituyen.

La respuesta arquitectónica

La brecha 2 —el acceso a datos explotado en cinco de las seis vulnerabilidades— tiene una respuesta arquitectónica directa. Kiteworks Compliant AI se sitúa entre los agentes de IA y los datos regulados, aplicando gobernanza en la capa de datos donde el modelo de IA no puede eludirla. Cada interacción pasa por cuatro puntos de control, independientes del modelo, prompt o framework del agente:

Identidad autenticada. Cada agente verificado mediante OAuth 2.0, vinculado a la persona autorizadora. Las credenciales se almacenan en el llavero del sistema operativo —nunca se exponen al modelo de IA. El control que faltó en el ataque al plugin de OpenAI.

Acceso controlado por políticas. Cada solicitud se evalúa en tiempo real según la identidad del agente, la clasificación de los datos y el contexto usando ABAC. Se aplica el acceso mínimo necesario a nivel de operación —no de sesión—. El control que faltó en EchoLeak, GeminiJack y ForcedLeak.

Cifrado validado FIPS 140-3. Todos los datos a los que accede el agente se cifran en tránsito y en reposo con módulos criptográficos validados.

Registro de auditoría a prueba de manipulaciones. Cada interacción se registra con atribución completa y se transmite en tiempo real al SIEM. Es la evidencia que exigen los reguladores —y que las seis vulnerabilidades no dejaron.

Los principios arquitectónicos que hacen efectivos estos controles —independencia del modelo de IA, aplicación por debajo de la capa de prompt, aislamiento de credenciales fuera del contexto de la IA— también cubren las brechas uno y tres. La implementación MCP de Kiteworks aplica políticas por operación, valida la navegación de rutas y mantiene las credenciales fuera del alcance del modelo. El framework está diseñado para evolucionar junto a las amenazas.

Tres preguntas para saber dónde estás parado

Brecha 1 (Confianza en la entrada): ¿Validas los datos externos antes de que la IA los procese? Si no, todas las vulnerabilidades de esta serie podrían funcionar en tu entorno.

Brecha 2 (Acceso a datos): ¿Cada solicitud de datos de IA se evalúa contra la política de forma independiente —en cada operación, no solo a nivel de sesión? Si no, una sola instrucción inyectada puede barrer todo tu alcance de datos.

Brecha 3 (Contención): ¿Sabes qué APIs y canales de salida pueden invocar tus procesos de IA internos? Si no, solo has evaluado el acceso a datos, pero no el alcance funcional. Combinado con entrada no confiable, eso es GrafanaGhost.

Más información